

00/61/07
490 U.S. PTO

07-21-A

A

Please type a plus sign (+) inside this box [+]

PTO/SB/05 (12/97)

Approved for use through 09/30/00. OMB 0651-0032

Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PC 09/619025 PTO
07/19/00

07/19/00

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No. 03760.P001X Total Pages 4, including Fee Transmittal

First Named Inventor or Application Identifier Garnet R. Chaney, et al.

Express Mail Label No. EL 14356 6555 US

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, D. C. 20231

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. Fee Transmittal Form
(Submit an original, and a duplicate for fee processing)
2. Specification (Total Pages 29, including cover page)
(preferred arrangement set forth below)
 - Descriptive Title of the Invention
 - Cross References to Related Applications
 - Statement Regarding Fed sponsored R & D
 - Reference to Microfiche Appendix
 - Background of the Invention
 - Brief Summary of the Invention
 - Brief Description of the Drawings (if filed)
 - Detailed Description
 - Claims
 - Abstract of the Disclosure
3. Drawings(s) (35 USC 113) (Total Sheets 11 sheets of 11 figures)
4. Oath or Declaration (Total Pages 5)
 - a. Newly Executed Original (**Partially executed**)
 - b. Copy from a Prior Application (37 CFR 1.63(d))
(for Continuation/Divisional with Box 17 completed) (**Note Box 5 below**)
 - i. DELETIONS OF INVENTOR(S) Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. Incorporation By Reference (useable if Box 4b is checked)
The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. Microfiche Computer Program (Appendix)

7. _____ Nucleotide and/or Amino Acid Sequence Submission
(if applicable, all necessary)

 - a. _____ Computer Readable Copy
 - b. _____ Paper Copy (identical to computer copy)
 - c. _____ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. Assignment Papers (cover sheet & documents(s))
9. a. 37 CFR 3.73(b) Statement (where there is an assignee)
 b. Power of Attorney
10. English Translation Document (if applicable)
11. a. Information Disclosure Statement (IDS)/PTO-1449
 b. Copies of IDS Citations
12. Preliminary Amendment
13. Return Receipt Postcard (MPEP 503) (Should be specifically itemized)
14. a. Small Entity Statement(s)
 b. Statement filed in prior application, Status still proper and desired
15. Certified Copy of Priority Document(s) (if foreign priority is claimed)
16. Other: 3 checks in the amounts of \$690.00, \$330.00, and \$40.00

17. If a **CONTINUING APPLICATION**, check appropriate box and supply the requisite information:

Continuation Divisional Continuation-in-part (CIP)
of prior application No: 09/165,675

18. Correspondence Address

Customer Number or Bar Code Label

(Insert Customer No. or Attach Bar Code Label here)

or

Correspondence Address Below

NAME James H. Salter (Reg. No. 35,668)

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP

ADDRESS 12400 Wilshire Boulevard

Seventh Floor

CITY Los Angeles

STATE California

ZIP CODE 90025-1026

Country U.S.A.

TELEPHONE (408) 720-8300

FAX (408) 720-9397

UNITED STATES PATENT APPLICATION

FOR

METHOD AND APPARATUS FOR GENERATING A LANGUAGE INDEPENDENT DOCUMENT ABSTRACT

INVENTORS:

GARNET R. CHANEY
ROBERT F. RICHARDSON
SEYMOUR I. RUBINSTEIN

PREPARED BY:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CA 90025-1026
(408) 720-8300

EXPRESS MAIL CERTIFICATE OF MAILING

"Express Mail" mailing label number: EL 14356 6555 US

Date of Deposit: July 19, 2000

I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee has been addressed to BOX PATENT APPLICATION, Assistant Commissioner for Patents, Washington, D. C. 20231.

Barbara Skliba

(Typed or printed name of person mailing paper or fee)

B2B2C Sklub

(Signature of person mailing paper or fee)

7/19/0

METHOD AND APPARATUS FOR GENERATING A LANGUAGE INDEPENDENT DOCUMENT ABSTRACT

CROSS-REFERENCE TO RELATED APPLICATION

5 This application claims the benefit of the filing date of U.S. Application No. 09/165,675, filed September 28, 1998, entitled "Language Independent Phrase Extraction".

FIELD OF THE INVENTION

The present invention relates to the field of computerized linguistic analysis and document management. More specifically, the present invention relates to automatic language independent identification of significant sentences in a machine-readable document.

BACKGROUND OF THE INVENTION

To ease the burden of sifting through the enormous volume of electronically available information, modern computer systems and other machines are often used to extract meaningful content from stores of information and to organize the content for a human operator. Many information dispensing services employ some sort of language analyzer for this purpose.

Machine-implemented language analyzers are usually one of two general types: referential analyzers and mathematical analyzers. Referential analyzers (also called semantic analyzers) typically use a combination of syntactic analysis and definitional analysis to identify significant phrases in a document. Syntactic analysis is used to parse paragraphs, sentences or other sequence of words into phrases and to remove conceptually insignificant terms, such as conjunctions, articles and prepositions. Definitional analysis involves identifying significant phrases by reference to the dictionary definitions of the terms constituting each phrase.

25 Typically, numeric weights are assigned to the words in a phrase according to their definitional

significance, and the average, sum or some other combination of the weights is used to represent the definitional significance of the phrase. Because the definitional significance of the phrase is expressed as a numeric value (sometimes called a “relevance code”), numeric thresholds can often be used to discriminate between significant and insignificant phrases according to application needs.

Referential analyzers suffer from a number of disadvantages, due mostly to their reliance on the definitional significance of words. First, some sort of database of words and their respective numeric weights is usually required. The database consumes memory and makes for relatively slow linguistic analysis because a separate database search is usually required for each word in a phrase. Another disadvantage of referential analyzers is that they are language dependent, requiring a different database of words for each language as well as specialized databases for different industries and fields. This places a significant burden on developers of referential analyzers and limits the applicability of systems that incorporate referential analyzers to the particular languages for which word databases are provided.

Mathematical analyzers perform linguistic analysis by measuring the relative frequency of occurrence of stemmed words. A stemmed word is a word that has been reduced to its root form by removing inflectional elements (e.g., indications of plurality, tense, case and so forth) and by truncating declensional and conjugative forms of the word. Groups of stemmed words having a relatively high frequency of occurrence relative to other stemmed words are considered to be significant phrases.

Unfortunately, mathematical analyzers suffer from many of the same disadvantages as referential analyzers. A database of stemmed words and their various inflected forms is usually required. As with referential analyzers, the database consumes memory and makes for relatively slow linguistic analysis because a separate database search is usually required for each word in a

phrase to determine whether there is a corresponding stemmed word. Mathematical analyzers are also language dependent and require a different database of words for each different language. As with referential analyzers, the language dependence of mathematical analyzers places a significant burden on developers of mathematical analyzers and limits applicability of systems that incorporate mathematical analyzers to the particular languages for which stemmed word databases are provided.

SUMMARY OF THE INVENTION

A method of extracting significant phrases from one or more documents stored in a computer-readable medium is disclosed. A sequence of words is read from the one or more documents and a score is determined for each word in the sequence based on the length of the word. The score for each word in the sequence is compared against a threshold score. The sequence of words is indicated to be a significant phrase if the number of words in the sequences that have a score greater than the threshold score equals or exceeds a predetermined number.

A sentence containing the sequence of words is retrieved from the document, if the sequence of words is a significant phrase. An abstract of the document is searched to determine if the sentence has been previously included in the abstract. If not, the sentence is added to the abstract.

Other features and advantages of the invention will be apparent from the accompanying drawings and from the detailed description that follows below.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements and in which:

5 Figure 1 is a high-level flow diagram of a method according to an embodiment of the present invention;

Figure 2 is a more detailed flow diagram of the method of Fig. 1 according to one embodiment;

Figure 3 is a flow diagram of a method for determining the score of a word according to one embodiment;

10 Figure 4 is a flow diagram of a method for conditionally adding a word to a phrase buffer according to one embodiment;

Figure 5 is a flow diagram of a method for conditionally adding a phrase to a list of phrases according to one embodiment;

15 Figure 6 is a diagram illustrating inputs and outputs of a phrase extraction engine according to one embodiment;

Figure 7 depicts a display of an application program in which an embodiment of the above-described phrase extraction engine is used;

Figure 8 depicts a display of the application program described in reference to Fig. 7 after a document called “creme.txt” has been processes by the phrase extraction engine;

20 Figure 9 depicts a display of the application program described in reference to Figs. 7 and 8 after the user has clicked on an Options menu and has selected a Phrase Verbosity control;

Figure 10 depicts a display of the application program described in reference to Figs. 7 & 8 and 9 after the user has selected a higher verbosity level and caused the “creme.txt” to be reprocessed by the phrase extraction engine; and

Figure 11 is a block diagram of a processing system that can be used to perform processing operations used in embodiments of the present invention.

DETAILED DESCRIPTION

According to embodiments disclosed herein, automatic phrase extraction is performed by analyzing a sequence of words and identifying sequences of relatively long or otherwise significant words to be significant phrases. Because phrase extraction is performed without regard to word definitions and without the need to reduce words to their stemmed form, phrase extraction according to embodiments described herein can be performed significantly faster and with less system resources than the above-described prior-art techniques. Further, because dictionary definitions and language-specific stemming information is not required to extract phrases from a document, the methods and apparatuses disclosed herein may be used to extract phrases in virtually any alphabet-oriented language without knowledge of the language beyond the particular punctuation marks and other symbols used to delimit words and groups of words from one another.

Figure 1 is a high-level flow diagram of a method for extracting significant phrases from a document stored in a machine-readable medium. At block 110, words are read from the document in sequence until a phrase delimiter is reached. Herein, the expression “phrase delimiter” refers to a period, question mark, exclamation point, comma, parenthesis, end-of-paragraph, sequence of non-alphabetic characters or other indication of the end of a sequence of words. As discussed below, overly long words and successive instances of relatively short words may also be treated as phrase delimiters. Herein, the term “word” refers to a string of characters (i.e., alphabet characters, numerals, quotation marks, hyphens and inflective punctuation marks such as apostrophes) that is separated from at least one other word by a word delimiter. A word delimiter may be a space, an end-of-line indicator, a paragraph break and other punctuation characters.

As discussed below, at least one embodiment of the linguistic analysis engine includes an interface to receive additional symbols that constitute word and phrase delimiters. This interface can be used to tailor the operation of the linguistic analysis engine according to application needs.

5 At block 120, a score is assigned to each word in the sequence of words based on the length of the word. As discussed below, the score may be increased if certain factors are present. For example, the score of a word may be increased by a predetermined amount for each letter in the word that is capitalized. Also, well known abbreviations and acronyms may be assigned increased scores. At block 130, the score for each word is compared against a threshold score. At block 140, the number of words having a score greater than the threshold score is determined and compared against a threshold number. If the number of words having a score greater than the threshold score exceeds the threshold number, the sequence of words is considered to be a significant phrase.

Figure 2 is a more detailed flow diagram of the method of Fig. 1 according to one embodiment. At block 201, a document pointer is set to the start of a document and a data structure for holding a list of significant phrases is initialized to an empty list. An average word score is also initialized to a predetermined stating number (e.g., five).

At block 203, a phrase buffer, a phrase word count and a qualified word count are each cleared to prepare to accumulate a new phrase. The phrase buffer, as its name implies, is used to 20 accumulate phrases that are conditionally added to the phrase list. The phrase word count is used to determine the total number of words in the phrase buffer and the qualified word count is used to determine the total number of qualified words in a phrase buffer. The purpose of the qualified word count is to indicate the extent to which insignificant words have been appended to a phrase. This operation is made clear in the following description.

At decision block 205, execution of the method proceeds to block 223 if the end of the document is detected and to block 207 if the end of the document is not detected. At block 207, a token is read from the document at the position pointed at by the document pointer and the pointer is advanced to point to the next token. Herein, a token refers to either a word or a phrase delimiter so that in each read operation in block 207, either the next word in a sequence of words or a phrase delimiter (e.g., a period, comma, question mark, etc.) is read. The token evaluated at block 209 to determine whether it is a phrase delimiter. If the token is a phrase delimiter, execution proceeds to block 215. If the token is not a phrase delimiter, the token is considered to be the current word and execution proceeds to block decision block 213. At decision block 213, the current word is inspected to determine if it satisfies some other phrase delimiting criteria. For example, in one embodiment, a word longer than a predetermined length (e.g., 40 characters) is treated as a phrase delimiter. Also, if a word is purely numeric and longer than a predetermined length, it may be treated as a phrase delimiter. If the word does satisfy a phrase delimiting criteria in block 213, execution proceeds to block 215, otherwise execution proceeds to block 217. At block 217, the score of the current word is calculated. According to one embodiment, the score of a word is equal to the number of characters in the word plus an increment (by one, for example) for each capitalized letter in the word. Word scores may be incremented for other particularities without departing from the scope of the present invention. At block 219, the score of the current word is used to update a running average of word scores.

5 In one embodiment, the initial average word score set in block 201 is weighted by a predetermined number (e.g., 4) so as to have a stabilizing effect on the average computed for the first few word scores. For example, if the average word score had been initialized to five in block 201, and then word scores of nine, eight, one, and seven had been computed in successive executions of block 217, the running average of word scores computed in block 219 would be:

10

$$\text{Running Average} = ((4*5)+9+8+1+7)/(4+4)=5.625$$

Also, in at least one embodiment, the running average is not permitted to fall below a predetermined minimum (e.g., 4). This value may be modified by a user.

At block 221, the current word is added to the accumulated phrase buffer if appropriate.

- 5 The logic of block 221 is described below in reference to Fig. 4. From block 221, execution loops back to decision block 205 to determine if the end of the document has been reached and, if not, to read and process the next token read from the document. If the end of the document is detected at block 205, block 215 is executed to determine whether to add contents of the phrase buffer to the list of phrases. The logic of block 215 is described in further detail below in reference to Fig. 5.

At block 220, a sentence containing the phrase is retrieved from the document, if the phrase is determined to be added to the phrase list. In one embodiment, the document is searched and the sentence containing the phrase is located. If the execution of block 215 determines that the contents of the phrase buffer should be added to the list of phrases, then the sentence is retrieved. At decision block 222, an abstract of the document is searched to determine if the retrieved sentence is included in the abstract. In one embodiment, the abstract consists of concatenated sentences that contain phrases being added to the phrase list. If the sentence is included in the abstract, then at block 225, the phrase list is returned to the entity that invoked the linguistic analysis. Otherwise, if the sentence is not included in the abstract, at block 20 224, the sentence is added to the abstract. In one embodiment, the sentence is concatenated to the other sentences within the abstract.

Next, at block 225, the list of phrases is returned and the process ends.

As discussed above, if the token is determined to be a phrase delimiter in block 209 or if the token satisfies a phrase delimiting criteria in block 213, execution proceeds to block 215.

Similar to the above discussion, at block 220, a sentence containing the phrase is retrieved from the document, if the phrase is determined to be added to the phrase list. In one embodiment, the document is searched and the sentence containing the phrase is located. If the execution of block 215 determines that the contents of the phrase buffer should be added to the
5 list of phrases, then the sentence is retrieved. At decision block 222, an abstract of the document is searched to determine if the retrieved sentence is included in the abstract. In one embodiment, the abstract consists of concatenated sentences that contain phrases being added to the phrase list. If the sentence is included in the abstract, then at block 225, the phrase list is returned to the entity that invoked the linguistic analysis. Otherwise, if the sentence is not included in the abstract, at block 224, the sentence is added to the abstract. In one embodiment, the sentence is concatenated to the other sentences within the abstract.

Next, at block 225, the list of phrases is returned and the process ends. Execution then loops back to block 205 to prepare to accumulate the next phrase in the document. As mentioned, the logic of block 219 is described in further detail below in reference to Fig. 5.
10
15

In an alternate embodiment, subsequent to the execution of block 215, if the phrase within the phrase buffer is determined to be added to the phrase list, and the phrase list is returned at block 225, each phrase within the phrase list may be retrieved and blocks 220, 222, and 224 may be executed to form the abstract.

In yet another alternate embodiment, blocks 201 through 225 may be executed after a
20 phrase list has been returned at block 225. In this embodiment, different abstracts may be created, based on the number of phrases selected as significant and stored within the phrase list.

Fig. 3 illustrates the logic of block 217 of fig. 2 in greater detail. At decision block 301, a table of fixed-length words is searched to determine if the current word is present in the table. As mentioned above, this operation is useful to prevent a word such as abbreviation or acronym

from being interpreted to be an insignificant word. As discussed below, an application programming interface may be used to provide a pointer to a table (or list) of fixed-length words so that operation of the phrase extraction engine described herein can be tailored according to user or application needs.

5 Returning to Fig. 3, if the current word is found in the table of fixed-length words, then a corresponding score for the current word is retrieved from the table at block 309 and used as the score of the word. If the current word is not found in the table of fixed-length words, then the length of the word is determined at block 303 and the number of capitalized letters in the current word is determined at block 305. The score for the current word is then determined based as a function of the length of the current word and the number of capitalized letters. For example, in one embodiment, the score is the sum of the word length and the number of capitalized letters. Other functions for determining the score based on the word length, the number of capitalized letters, or both may be used without departing from the scope of the present inventions. After the score is determined at block 307 or retrieved at block 309, execution returns to the calling block.

Fig. 4 illustrates the logic of block 221 of Fig. 2 in greater detail. At decision block 401, the score of the current word is compared to a threshold score. In one embodiment, the threshold score is the running average score that was updated in block 219 of Fig. 2 plus an adjustable delta value. If the score of the current word exceeds the threshold score, then the qualified word count is set to the phrase word count plus one in block 421. The current word is then added to the phrase buffer 417 and the phrase word count is incremented in block 419. Each new word added to the phrase buffer in block 417 is appended to the group of words already in the phrase buffer to build up the phrase. After block 419, the method returns to the calling block in Fig. 2.

Returning to decision block 401, if the threshold score is not exceeded, then the word is considered to be a “short” word and execution proceeds to decision block 403. At decision block 403, the qualified word count is evaluated to determine whether the phrase buffer contains any qualified words.. If not, the word is not added to the phrase buffer and execution returns to the 5 calling block in Fig. 2. The effect of the operation is to ignore short words that begin a phrase.

If, at decision block 403, the qualified word count is determined to be a greater than zero (i.e., there is at least one qualified word in the phrase buffer), then at decision block 407, the phrase buffer is inspected to determine whether the phrase buffer contains any short words. In one embodiment, a Boolean variable is set when a short word is added to the phrase buffer and cleared when the phrase buffer is cleared. If there are no short words in the phrase buffer, then execution proceeds to block 417 where the current (short) word is entered into the phrase buffer and then to block 419 to increment the phrase word count as described above. In the embodiment that includes the Boolean variable that is set when a short word is added to the phrase buffer, then Boolean variable would be set when the short word is added to the phrase buffer at block 417.

Returning to decision block 407, if there is a short word (or short words) in the phrase buffer, then at decision block 411 the word most recently added to the phrase buffer is inspected to determine whether it is a short word. If not, execution proceeds to block 215 to conditionally add the phrase to the list, after which execution of the method returns to the calling block in Fig. 20 2. The logic of block 215 is described in further detail below in reference to Fig. 5.

If the word most recently added to the phrase buffer is determined to be a short word, then the phrase buffer is inspected at block 415 to determine if the last N words added to the phrase buffer are also short words. If so, then block 215 is entered to conditionally add the

phrase to the list. If the last N words added to the phrase buffer are not short words, then blocks 417 and 419 are entered as described above before returning to the calling block in Fig. 2.

5

Reflecting on the above described operation for determining whether to add the current word to the phrase buffer, it will be noted that, in the embodiment described, sequences of N short words are treated as phrase delimiters (see block 415). Also, by the operation of block 411, a phrase generally will not have more than a single sequence (i.e., one or more) of short words embedded between qualified words. In alternate embodiments, the number of separate sequences of embedded of short words may be increased. Further, referring to the operation at block 421, the purpose of the assignment of the phrase word count plus 1 to the qualified word count should be clear: the qualified word count tracks the phrase word count except for short words added to the end of a phrase. Consequently, if a phrase ends with one or more short words, the qualified word count will not reflect the addition of the short words to the phrase buffer and thus the qualified word count can be used to limit the phrase to the qualified words and any short words embedded between qualified words.

15

20

Fig. 5 illustrates the logic of block 215 of Figs. 2 and 4 in greater detail. At block 501, the phrase word count is set to the qualified word count. The effect of this operation is to effectively remove any short words from the end of phrase in the phrase buffer. The phrase word count (or qualified word count) is then compared against a threshold value (e.g., a minimum phrase length) at decision block 503 to determine whether the phrase in the phrase buffer should be added to the phrase list. If the phrase word count exceeds the threshold, the phrase in the phrase buffer is added to the phrase list at 507 and the phrase buffer, qualified word count and phrase word count are cleared at block 511. Execution then returns to the calling block. If the phrase word count does not exceed the threshold value in block 503, execution returns to the calling block without adding the phrase to the phrase buffer.

In one embodiment, a phrase extraction engine that operates as described above is implemented by execution of a program of instructions in a processor. The program of instructions include software routines which can be called by application programs to initialize the phrase extraction engine and to control the output generated by the phrase extraction engine.

- 5 The software routines that can be called by application programs define an application programming interface (API) and, in one embodiment, includes routines that can be called to provide a pointer to a document to be processed by the phrase extraction engine; provide a verbosity setting to the phrase extraction engine to control the verbosity of phrases considered by the phrase extraction engine to be significant phrases; provide to a table of fixed length words that can be referenced by the phrase extraction engine; provide to the phrase extraction engine a list of characters (or sequence of characters) that are considered to be phrase delimiters; provide to the phrase extraction engine a list of characters (or sequences of characters) that are considered to be word delimiters; provide to the phrase extraction engine a list of characters (or sequences of characters) that constitute the alphanumeric character set used in a particular language and so forth.

10
15

20

Fig. 6 depicts a phrase extraction engine 620 that included an API for receiving a pointer to a document 601, a verbosity setting 605 and a table of fixed words 611. In one embodiment, a single verbosity setting is used to set the minimum length of a phrase that is considered by the phrase extraction engine 620 to be significant (e.g., in block 401 of Fig. 4) and the predetermined value that is added to the running average length of words in block 219 of Fig. 2 to set the threshold score. In an alternate embodiment, separate routines may be provided in the API to set these different parameters. Also, the verbosity setting may be user supplied or automatically determined. Moreover, because the number of significant phrases found in a given document decreases as the verbosity level goes up (i.e., as more selective phrase identification criteria are

used), the verbosity may be changed dynamically to scale the number of significant phrases in the output list of phrases 625 to fall within a user selected range or at a user selected target. As discussed above in reference to Fig. 3, the table of fixed length words 611 is used to set scores for certain words (e.g., abbreviations and acronyms).

Fig. 7 depicts a display 701 of an application program in which an embodiment of the above-described phrase extraction engine is used. The application program presents a “Search for:” prompt 703 to prompt a user to specify the location and name of a document or set of documents to be processed by the phrase extraction engine. A uniform resource locator may also be entered to direct the application program to feed a page on the World Wide Web (the “web”) to the phrase extraction engine. The application program is also capable of automatically identifying a set of documents to be processed by the phrase extraction engine. A user may enter one or more search criteria and the application program searches locally or remotely (e.g., on the web) for documents that meet the search criteria. The phrase extraction engine is then invoked to process the one or more documents in the document set and to generate a list of significant phrases found in the respective documents. Indexing logic in the application program can then be used to automatically generate an alphabetically organized cross-index of significant phrases found in the document set by the phrase extraction engine and to display the significant phrases for the user. The above-described phrase extraction engine may also be used to automatically generate an index for a document, a cross index for a set of documents (e.g., showing cross-references to phrases found in more than one document or otherwise indicating that a phrase is found in more than one document), automatically generating hypertext links for a document and so forth.

Fig. 8 depicts a display 801 of the application program described in reference to Fig. 7 after a document called “creme.txt” has been processed by the phrase extraction engine. In

response to the user selecting or clicking a “phrases” tab 803 a phrase window 805 that includes an alphabetized list of the significant phrases is displayed. In one embodiment, the phrase extraction engine automatically processes a document when the document is identified by the application program. In an alternate embodiment, the phrase extraction engine may not process a
5 document until the phrases tab is selected. The phrase window 805 includes an alphabetized set of tabs that can be used to quickly move to a point in the list of significant phrases. For example, if a large number of significant phrases was found in a document (or set of documents), there may be several screens worth of phrases for each letter of the alphabet. Consequently, by selecting a particular letter in the tab, the set of phrases (or a portion of the set of phrases) starting with the letter are displayed in the phrase window 805. The list of significant phrases may also be displayed in the phrase window 805 in a different order and according to a different criteria. For example, the list of significant phrases can be displayed in the phrase window 805 according to the number of words per phrase, with the phrase having the most words displayed at the start or end of the list. The list of significant phrases 412 could also be displayed according to the total score of the words that appear in a significant phrase.
10
15

Still referring to Fig. 8, a refine button 807 is provided to allow the user to further control the display of the list of significant phrases in the phrase window 805. When the user clicks the refine button, a small refine window is displayed showing two options: a set phrase refine option and a clear phrase refine option. If the user selects the set phrase refine option, a set refine
20 specification window is displayed to allow the user to enter a word, a phrase, or a Boolean expression in a specification box to narrow the list of significant phrases displayed according to the specification entered by the user. For example, if the user wants to see the significant phrases that contain the word “information” but not the word “informative”, the user would enter the following in the specification box: “information AND NOT informative”. The clear phrase

option can be selected by the user to restore the full list of significant phrases to the phrase window 805.

Fig. 9 depicts a display 901 of the application program described in reference to Figs. 7 and 8 after the user has clicked on an Options menu and has selected a Phrase Verbosity control.

5 As indicated, a verbosity selection menu is presented to prompt the user to select one of seven of verbosity settings that range from least phrases per page to most phrases per page. In an alternate embodiment, a different range of settings may be used. As shown in Fig. 9, the Options menu also includes an option for number filtering. The number filtering option can be used to prevent words that contain more than a user-specified number of digits from being considered by the phrase extraction engine to be qualified words (i.e., all words with more than the specified number of digits are considered to be “short” words as that term is defined above). The number filtering option can also be used to cause the phrase extraction engine to entirely ignore words that have more than the user-specified number of digits.

Fig. 10 depicts a display 1001 of the application program described in reference to Figs. 7, 8, and 9 after the user has selected a higher verbosity level (less phrases per document) and caused the “creme.txt” to be reprocessed by the phrase extraction engine. As indicated, the phrase window contains far fewer significant phrases as a result of the more selective criteria used by the phrase extraction engine to identify significant phrases.

Fig. 11 is a block diagram of a processing system 50 that can be used to perform
20 processing operations used in embodiments of the present invention. The processing system 50 includes a processing unit 51, memory 53, display device 55, user 0 input device 57, and communications device 59 each coupled to a bus structure 61. In one embodiment, the processing system 50 is a general purpose computer system such as a desktop or laptop computer or a workstation or larger computer. The user-input device 57 may be a mouse, trackball, stylus,

or any other device for manipulating elements displayed on display device 55. Moreover, user-input device 57 may include a keypad or keyboard to allow a user to input alphanumeric data into the processing system 51. Other I/O devices 63 may be present according to the specific functions performed by the processing system 50.

5 The processing unit 51 may include one or more general purpose processors, one or more digital signal processors or any other devices capable of executing a sequence of instructions. When programmed with appropriate instructions, the processing unit may be used to implement the above-described embodiments of a phrase extraction engine.

10 The communication device 59 may be a modem, area network card or any other device for coupling the processing system 50 to a computer network. The communications device may be used to generate or receive a signal that is propagated via a conductive or wireless media. The propagated signal may be used, for example, for obtaining pages of text, images or other documents from a server computer on the World Wide Web or other network, or for receiving updated program code or function-extending program code that can be executed by the processing unit to implement embodiments of the present invention.

15 The memory 53 may include both system memory and non-volatile storage devices such as magnetic tape, magnetic disk, optical disk, electrically erasable programmable read only memory (EEPROM), or any other computer-readable medium. As shown in Fig. 11, the memory 53 may be used to store program code 54 for performing the above-described processing operations required to implement a phrase extraction engine and to implement the above-described user-interface of an application program. In one embodiment, when power is applied to the processing system 50, operating system program code is loaded from non-volatile storage into system memory by the processing unit 51 or another device, such as a direct memory access controller (not shown). Sequences of instructions comprised by the operating

system are then executed by processing unit 51 to load other sequences of instructions, including the above-described application programs, from non-volatile storage into system memory. Thus, embodiments of the present invention may be implemented by obtaining sequences of instructions from a computer-readable medium, including the above-described propagated signal, and executing the sequences of instructions in the processing unit 51.

Having described a processing system for implementing embodiments of the present invention, it should be noted that the individual processing operations described above may also be performed by specific hardware components that contain hard-wired logic to carry out the recited operations or by any combination of programmed processing components and hard-wired logic. Nothing disclosed herein should be construed as limiting the present invention to a single embodiment wherein the recited operations are performed by a specific combination of hardware components.

CLAIMS

What is claimed is:

1 1. A method of identifying a significant phrase in a document, the method
2 comprising:
3 reading a sequence of words from the document;
4 determining a score for each word in the sequence based on the length of each word;
5 comparing the score for each word in the sequence against a threshold score;
6 indicating that the sequence of words is a significant phrase if the number of words in the
7 sequence that have the score greater than the threshold score equals or exceeds a predetermined
8 number;

9 retrieving a sentence from the document, the sentence containing the sequence of words,
10 if the sequence of words is a significant phrase; and

11 searching an abstract of the document to determine whether the sentence is included in
12 the abstract.

1 2. The method of claim 1, wherein reading the sequence of words in the document
2 further comprises reading the sequence of words in the document until a phrase delimiter is
3 detected.

1 3. The method of claim 1, wherein the score for the word is proportional to the
2 number of characters in the word.

1 4. The method of claim 3, wherein the number of characters in the word includes the
2 number of numeric digits in the word.

1

1 5. The method of claim 1, wherein the score for the word is increased for each
2 capitalized letter in the word.

1

1 6. The method of claim 1, wherein the score for the word is the length of the word
2 plus the number of capitalized letters in the word.

1

7. The method of claim 1, wherein determining the score for each word in the sequence further includes:

determining whether the word exists in a predetermined table; and

retrieving the score for the word from the predetermined table.

8. The method of claim 1, wherein the threshold score is based on an average of the
2 scores of words in the sequence of words.

1

1 9. The method of claim 8, wherein the threshold score is adjusted by a

2 predetermined value to be different from the average of the scores of words in the sequence of
3 words.

1

1 10. The method according to claim 1, further comprising including the sentence in the
2 abstract, if the sentence is not included in the abstract.

1

1 11. The method according to claim 1, further comprising discarding the sentence if
2 the sentence is included in the abstract.

1

1 12. The method according to claim 1, wherein the abstract is language independent.

1

1 13. A method of identifying a significant phrase in a document, the method
2 comprising:

3 reading a sequence of words from the document;

4 determining a score for each word in the sequence based on the length of each word;

5 comparing the score for each word in the sequence against a threshold score;

6 indicating that the sequence of words is a significant phrase if the number of words in the
7 sequence that have the score greater than the threshold score equals or exceeds a predetermined
8 number;

9 storing the sequence of words and the number of words in the sequence, if the sequence
10 of words is a significant phrase.

1

1 14. The method according to claim 13, further comprising:

2 retrieving a sentence from the document, the sentence containing the sequence of words,

3 if the sequence of words is a significant phrase; and

4 searching an abstract of the document to determine whether the sentence is included in
5 the abstract.

1

1 15. The method according to claim 14, further comprising including the sentence in
2 the abstract, if the sentence is not included in the abstract.

1

1 16. The method according to claim 14, further comprising discarding the sentence if
2 the sentence is included in the abstract.

1

1 17. The method according to claim 14, wherein the abstract is language independent.

1

1 18. A computer readable medium containing executable instructions which, when
2 executed in a processing system, cause the system to perform a method for identifying a
3 significant phrase in a document, the method comprising:

4 reading a sequence of words from the document;

5 determining a score for each word in the sequence based on the length of each word;

6 comparing the score for each word in the sequence against a threshold score;

7 indicating that the sequence of words is a significant phrase if the number of words in the
8 sequence that have the score greater than the threshold score equals or exceeds a predetermined
9 number;

10 retrieving a sentence from the document, the sentence containing the sequence of words,

11 if the sequence of words is a significant phrase; and

12 searching an abstract of the document to determine whether the sentence is included in
13 the abstract.

1

1 19. The computer readable medium according to claim 18, wherein reading the
2 sequence of words in the document further comprises reading the sequence of words in the
3 document until a phrase delimiter is detected.

1

1 20. The computer readable medium according to claim 18, wherein the score for the
2 word is proportional to the number of characters in the word.

1

1 21. The computer readable medium according to claim 20, wherein the number of
2 characters in the word includes the number of numeric digits in the word.

1

1 22. The computer readable medium according to claim 18, wherein the score for the
2 word is increased for each capitalized letter in the word.

1

23. The computer readable medium according to claim 18, wherein the score for the
word is the length of the word plus the number of capitalized letters in the word.

1

24. The computer readable medium according to claim 18, wherein determining the
score for each word in the sequence further includes:

4 determining whether the word exists in a predetermined table; and
 retrieving the score for the word from the predetermined table.

1

1 25. The computer readable medium according to claim 18, wherein the threshold
2 score is based on an average of the scores of words in the sequence of words.

1

1 26. The computer readable medium according to claim 25, wherein the threshold
2 score is adjusted by a predetermined value to be different from the average of the scores of
3 words in the sequence of words.

1

1 27. The computer readable medium according to claim 18, further comprising
2 including the sentence in the abstract, if the sentence is not included in the abstract.

1
1 28. The computer readable medium according to claim 18, further comprising
2 discarding the sentence if the sentence is included in the abstract.

1
1 29. The computer readable medium according to claim 18, wherein the abstract is
2 language independent.

1
1
1 30. A computer readable medium containing executable instructions which, when
executed in a processing system, cause the system to perform a method for identifying a
significant phrase in a document, the method comprising:

1 reading a sequence of words from the document;

2 determining a score for each word in the sequence based on the length of each word;

3 comparing the score for each word in the sequence against a threshold score;

4 7 indicating that the sequence of words is a significant phrase if the number of words in the
5 sequence that have the score greater than the threshold score equals or exceeds a predetermined
6 number;

7
8 10 storing the sequence of words and the number of words in the sequence, if the sequence
9 of words is a significant phrase.

1
1 31. The computer readable medium according to claim 30, wherein the method
2 further comprises:

3 retrieving a sentence from the document, the sentence containing the sequence of words,
4 if the sequence of words is a significant phrase; and
5 searching an abstract of the document to determine whether the sentence is included in
6 the abstract.

1

1 32. The computer readable medium according to claim 31, wherein the method
2 further comprises including the sentence in the abstract, if the sentence is not included in the
3 abstract.

4

1 33. The computer readable medium according to claim 31, wherein the method
2 further comprises discarding the sentence if the sentence is included in the abstract.

3

1 34. The computer readable medium according to claim 31, wherein the abstract is
2 language independent.

1

ABSTRACT

A method of extracting significant phrases from one or more documents stored in a computer-readable medium. A sequence of words is read from the one or more documents and a score is determined for each word in the sequence based on the length of the word. The score for 5 each word in the sequence is compared against a threshold score. The sequence of words is indicated to be a significant phrase if the number of words in the sequences that have a score greater than the threshold score equals or exceeds a predetermined number. A sentence containing the sequence of words is retrieved from the document, if the sequence of words is a significant phrase. An abstract of the document is searched to determine if the sentence has been previously included in the abstract. If not, the sentence is added to the abstract.

SEARCHED
SERIALIZED
INDEXED
FILED

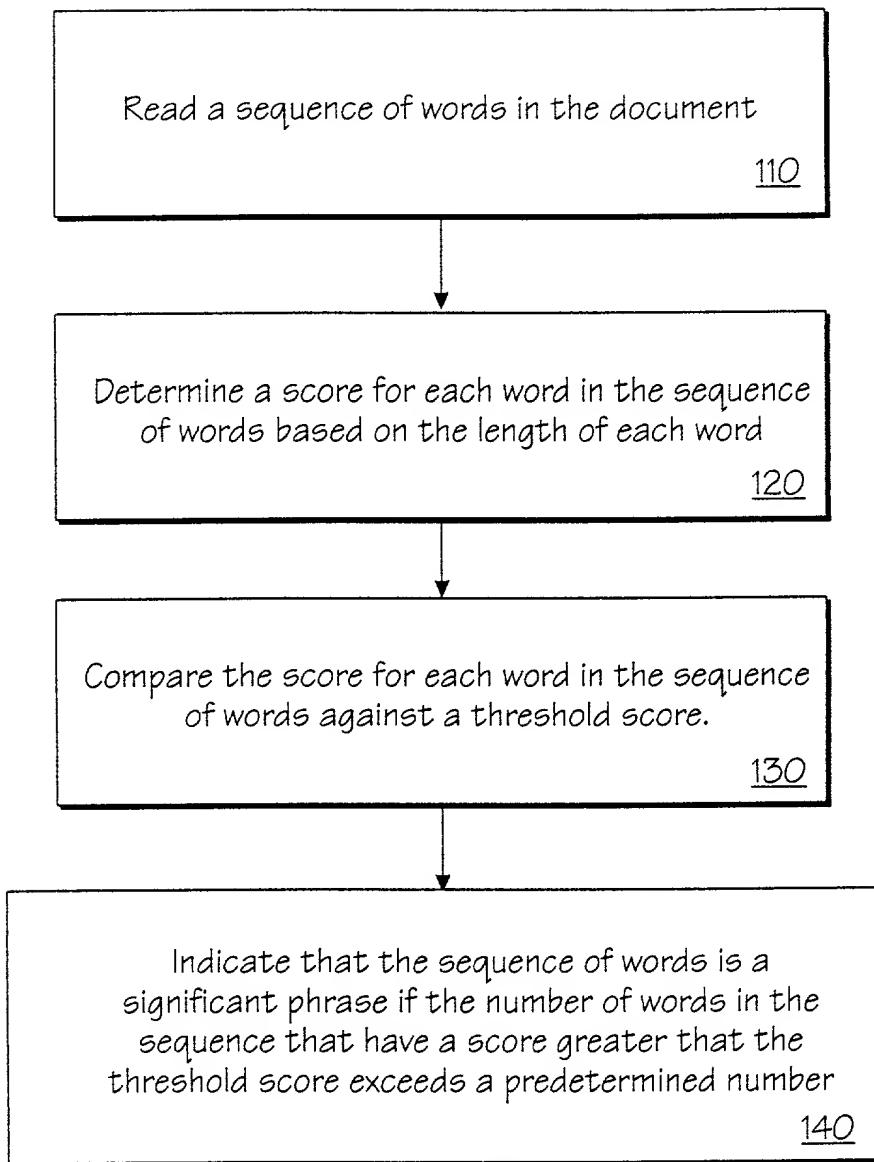


Fig. 1

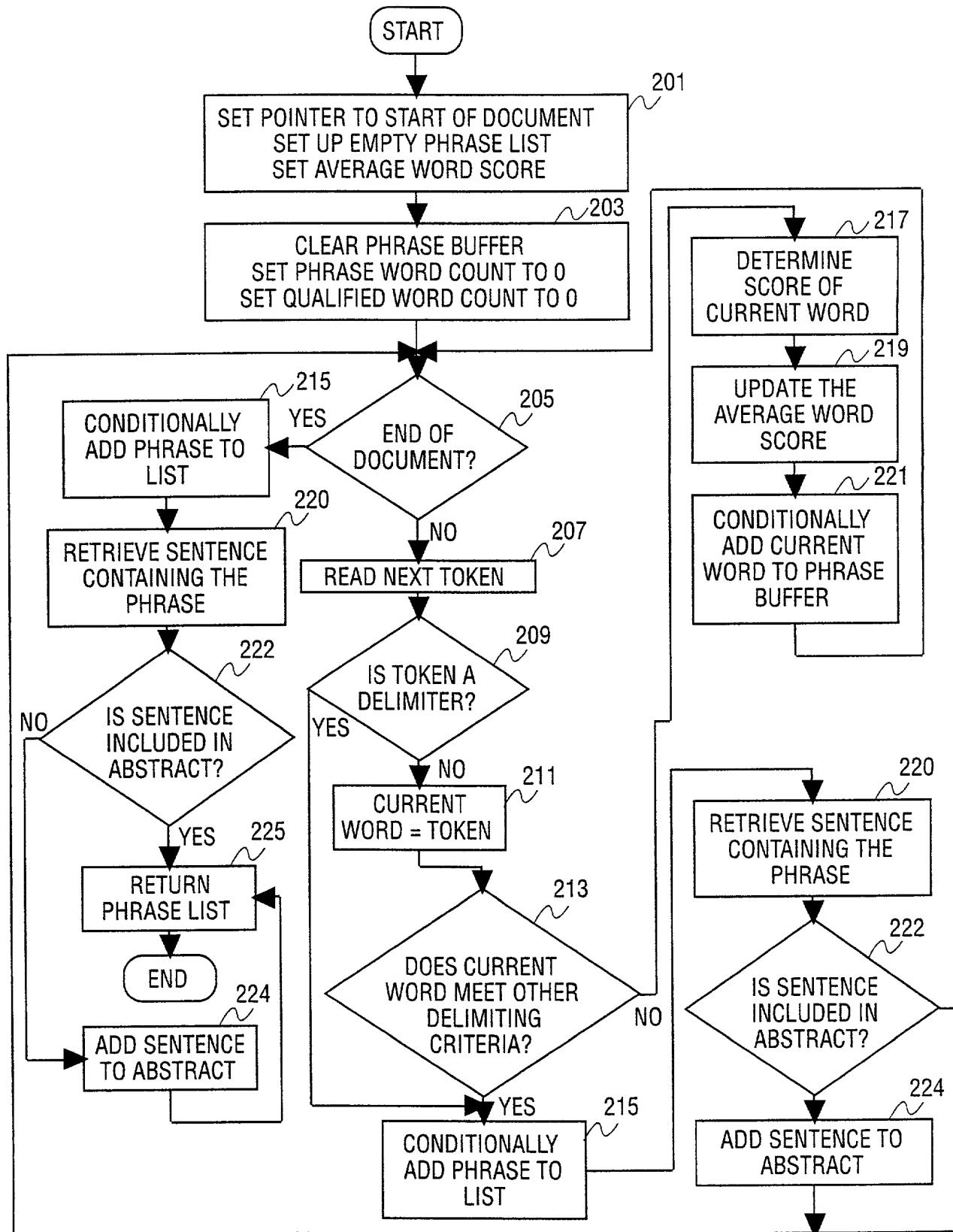


FIG. 2

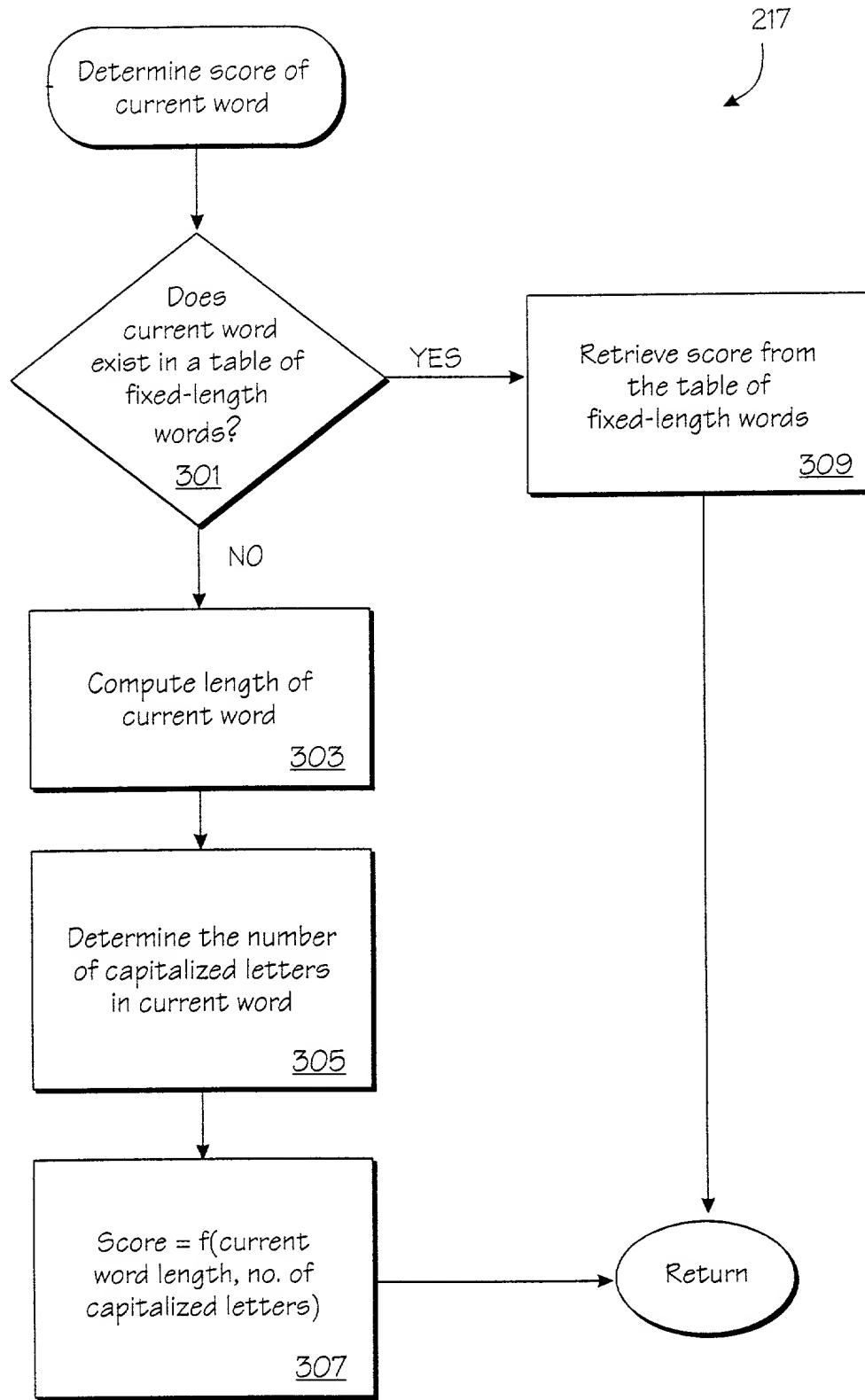


Fig. 3

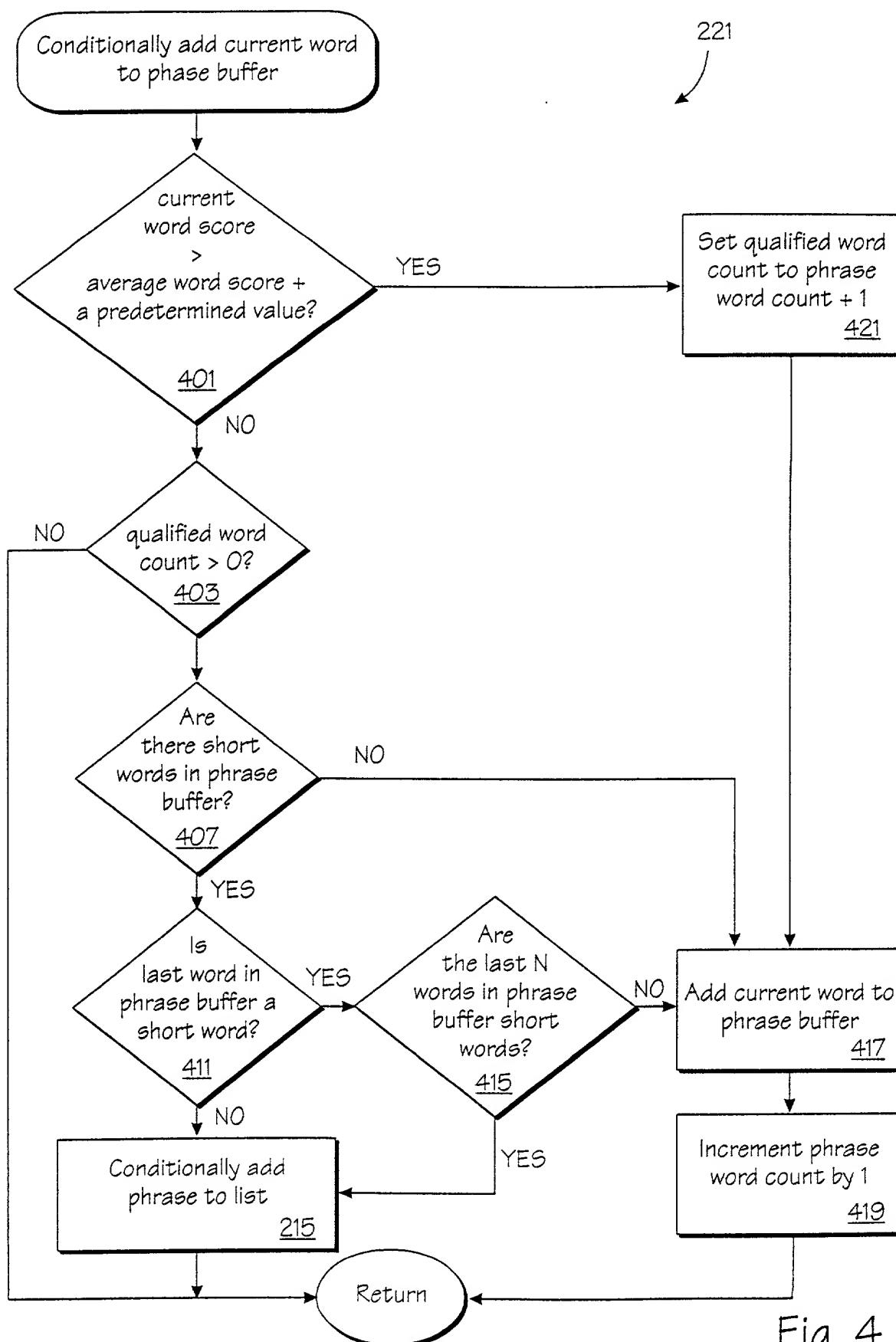


Fig. 4

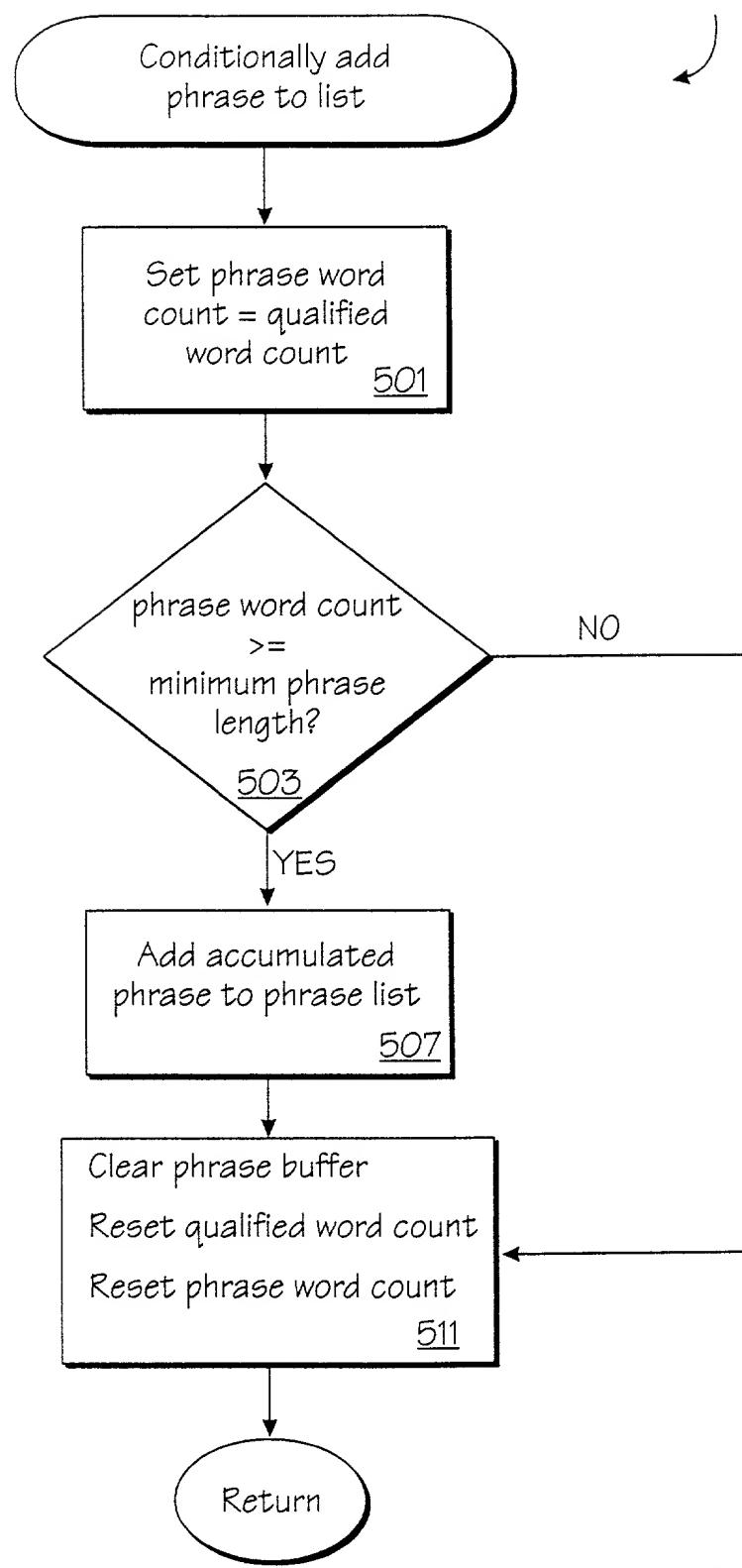


Fig. 5

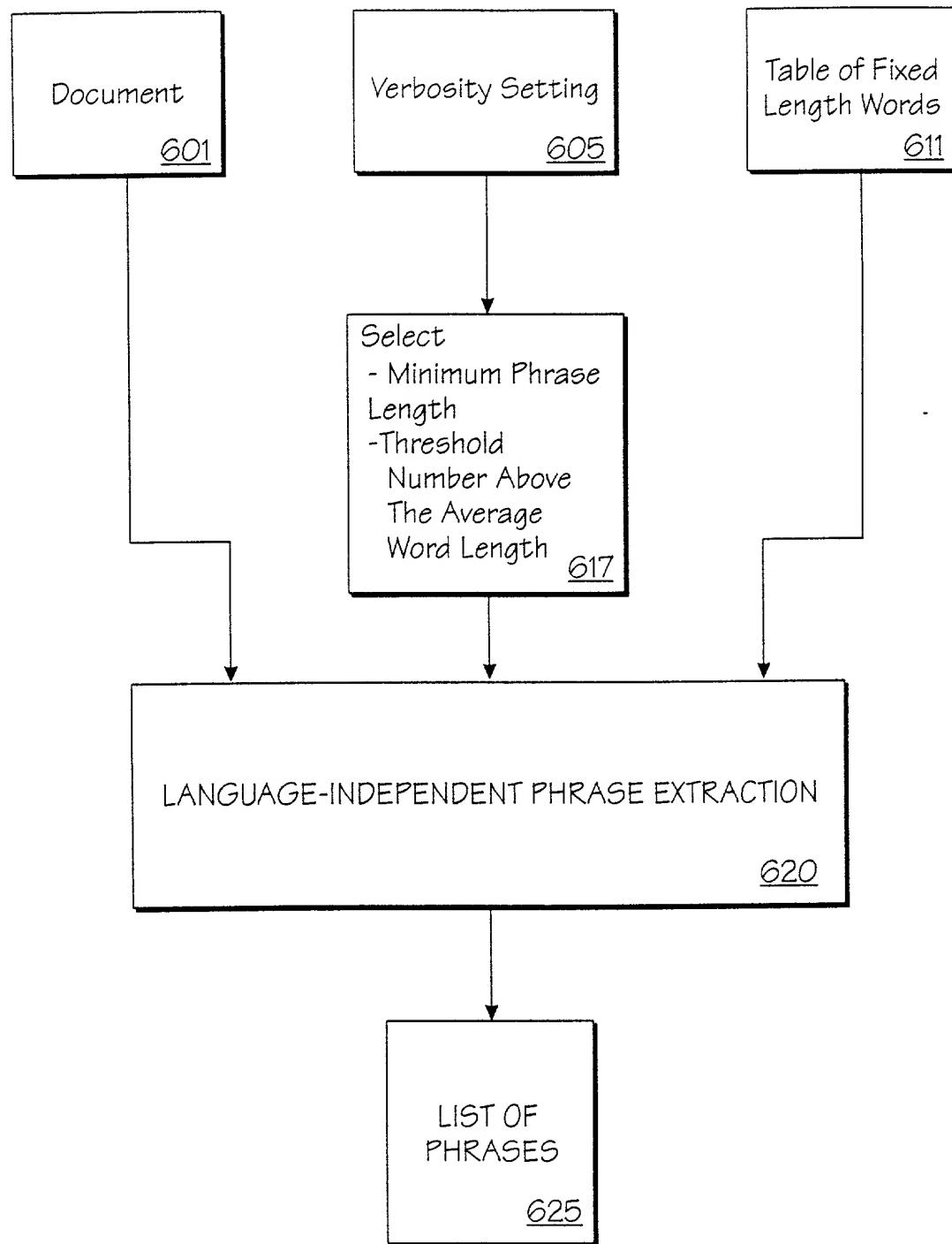


Fig. 6

Fig. 7

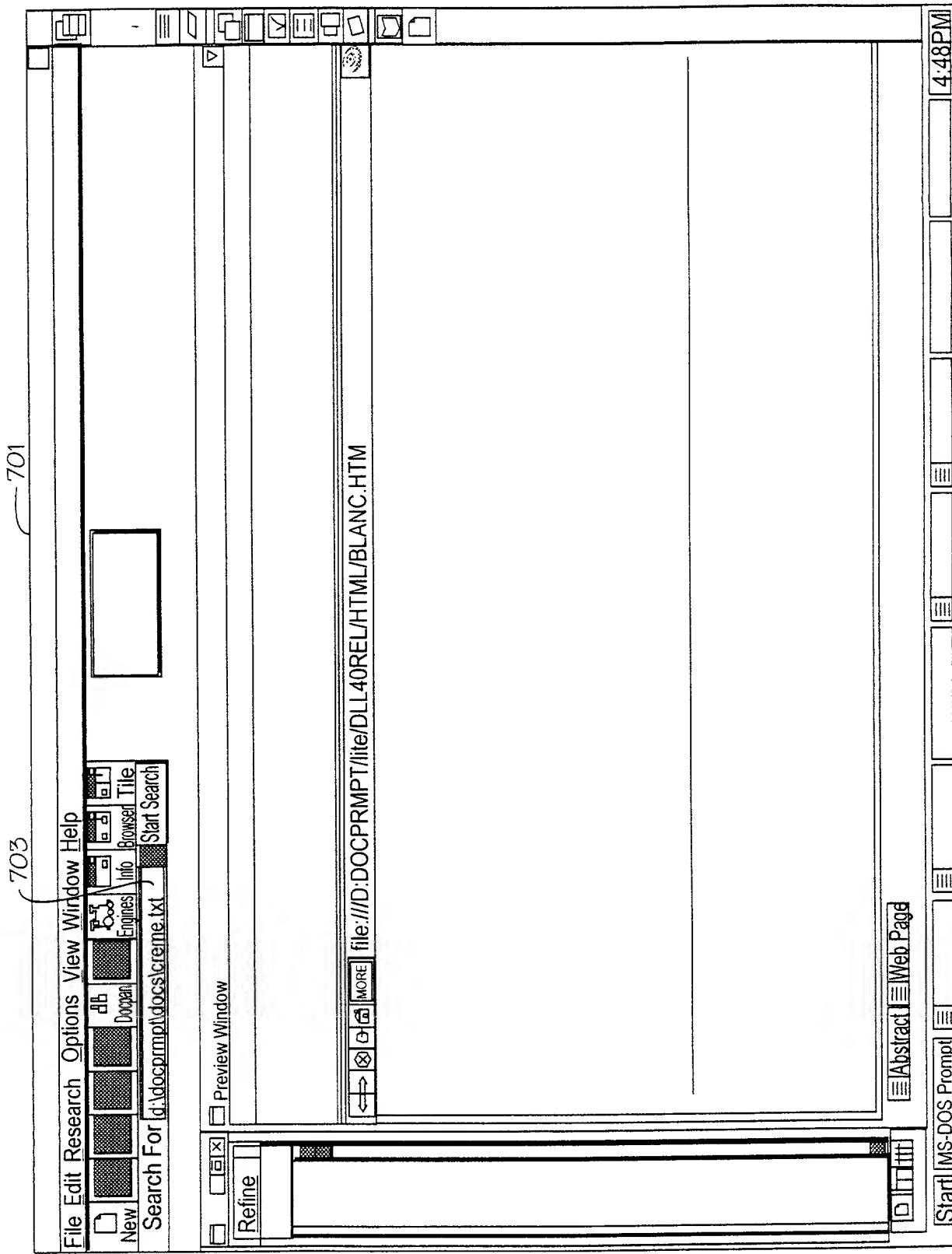


Fig. 8

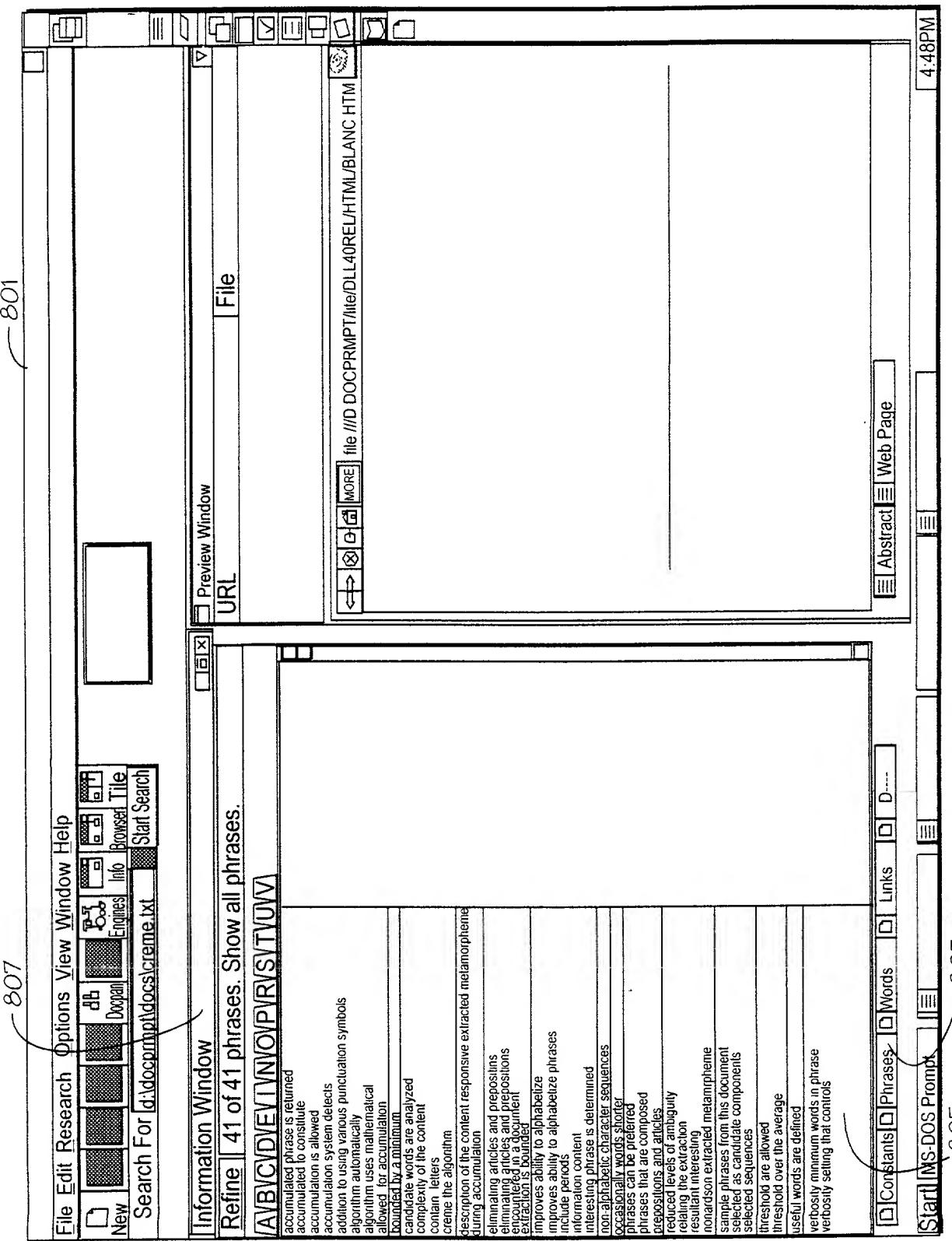


Fig. 9

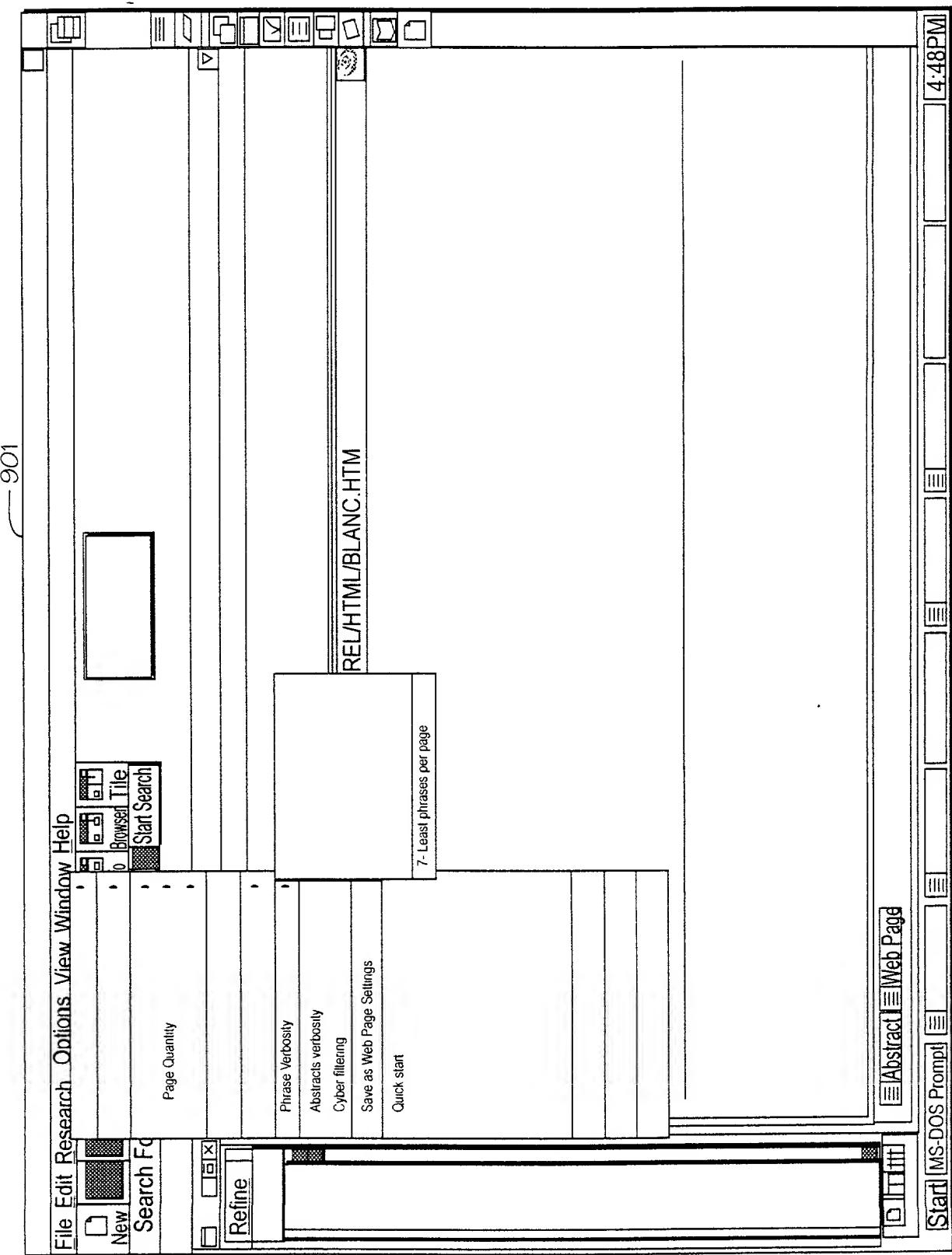


Fig. 10

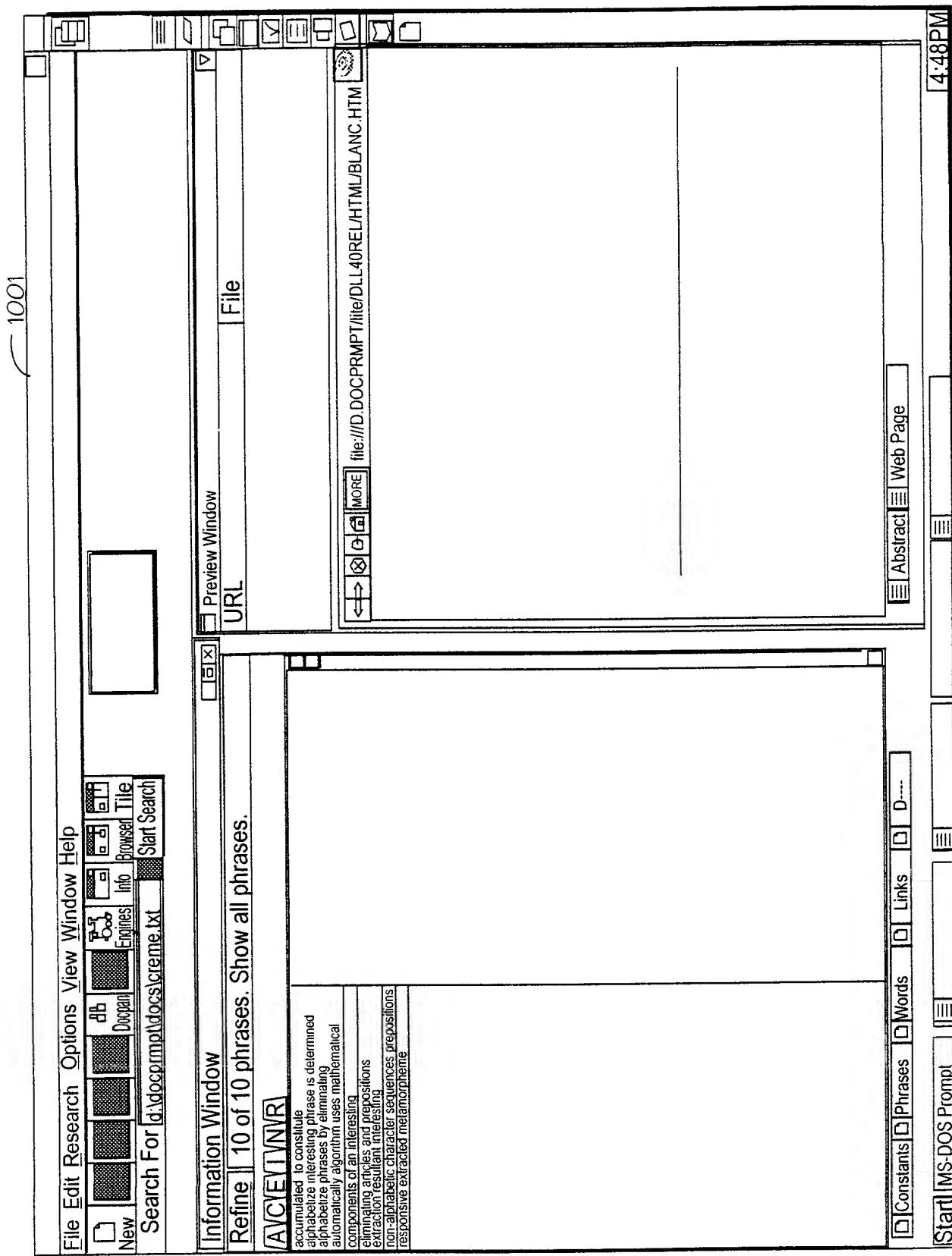
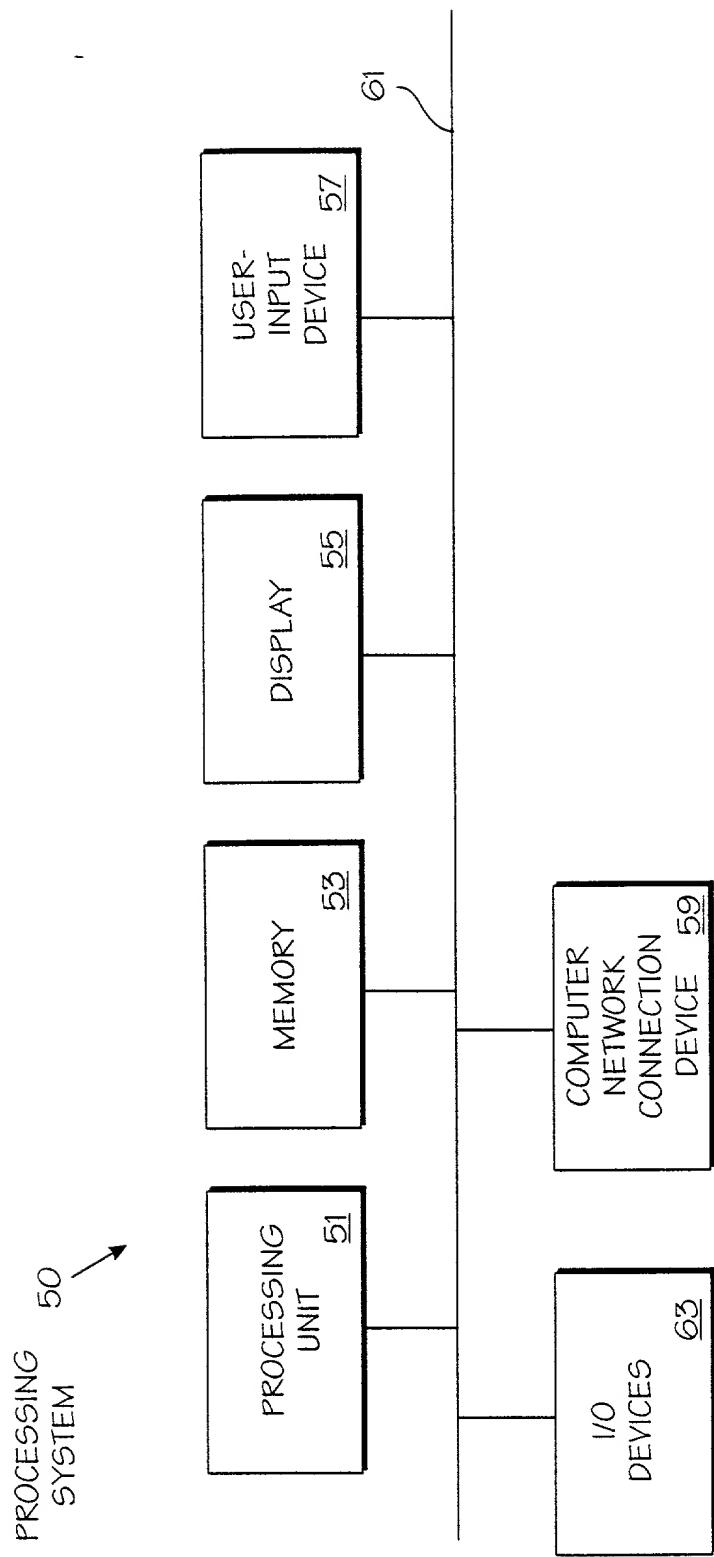


Fig. 11



**DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION
(CONTINUATION-IN-PART)**

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below, next to my name.

I believe I am the original, first, and sole inventor (if only one name is listed below) or an original, first, and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled Method And Apparatus For Generating A Language Independent Document Abstract

the specification of which

XX is attached hereto.
____ was filed on _____ as
United States Application Number _____
or PCT International Application Number _____
and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d), of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

<u>Prior Foreign Application(s)</u>			<u>Priority Claimed</u>	
<u>(Number)</u>	<u>(Country)</u>	<u>(Day/Month/Year Filed)</u>	<u>Yes</u>	<u>No</u>
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

I hereby claim the benefit under title 35, United States Code, Section 119(e) of any United States provisional application(s) listed below:

<u>Application Number</u>	<u>Filing Date</u>
_____	_____
I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each	

of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

<u>09/165,675</u>	<u>9/28/1998</u>	<u>Allowed</u>
<u>Application Number</u>	<u>Filing Date</u>	<u>Status -- patented, pending, abandoned</u>
<u>Application Number</u>	<u>Filing Date</u>	<u>Status -- patented, pending, abandoned</u>

I hereby appoint the persons listed on Appendix A hereto (which is incorporated by reference and a part of this document) as my respective patent attorneys and patent agents, with full power of substitution and revocation, to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith.

Send correspondence to James H. Salter, BLAKELY, SOKOLOFF, TAYLOR &
(Name of Attorney or Agent)
ZAFMAN LLP, 12400 Wilshire Boulevard 7th Floor, Los Angeles, California 90025 and direct telephone calls to James H. Salter, (408) 720-8300.
(Name of Attorney or Agent)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of Sole/First Inventor Garnet R. Chaney

Inventor's Signature _____ Date _____

Residence Novato, California Citizenship United States Of America
(City, State) (Country)

Post Office Address 167 San Marin Drive #101
Novato, California 94945

Full Name of Second/Joint Inventor Robert F. Richardson

Inventor's Signature Robert Richardson Date 17 July, 2000

Residence Novato, California Citizenship United States of America
(City, State) (Country)

Post Office Address 980 Ignacio Boulevard #15
Novato, California 94949

Full Name of Third/Joint Inventor Seymour I. Rubinstein

Inventor's Signature Seymour I. Rubinstein Date 17 July 2000

Residence Belvedere, California Citizenship United States Of America
(City, State) (Country)

Post Office Address 337 Belvedere Avenue
Belvedere, California 94920

APPENDIX A

William E. Alford, Reg. No. 37,764; Farzad E. Amini, Reg. No. P42,261; Aloysius T. C. AuYeung, Reg. No. 35,432; William Thomas Babbitt, Reg. No. 39,591; Carol F. Barry, Reg. No. 41,600; Jordan Michael Becker, Reg. No. 39,602; Lisa N. Benado, Reg. No. 39,995; Bradley J. Bereznak, Reg. No. 33,474; Michael A. Bernadicou, Reg. No. 35,934; Roger W. Blakely, Jr., Reg. No. 25,831; R. Alan Burnett, Reg. No. 46,149; Gregory D. Caldwell, Reg. No. 39,926; Andrew C. Chen, Reg. No. 43,544; Thomas M. Coester, Reg. No. 39,637; Donna Jo Coningsby, Reg. No. 41,684; Florin Corie, Reg. No. 46,244; Dennis M. deGuzman, Reg. No. 41,702; Stephen M. De Clerk, Reg. No. P46,503; Michael Anthony DeSanctis, Reg. No. 39,957; Daniel M. De Vos, Reg. No. 37,813; Robert Andrew Diehl, Reg. No. 40,992; Sanjeet Dutta, Reg. No. P46,145; Matthew C. Fagan, Reg. No. 37,542; Tarek N. Fahmi, Reg. No. 41,402; George Fountain, Reg. No. 37,374; Paramita Ghosh, Reg. No. 42,806; James Y. Go, Reg. No. 40,621; James A. Henry, Reg. No. 41,064; Libby N. Ho, Reg. No. P46,774; Willmore F. Holbrow III, Reg. No. P41,845; Sheryl Sue Holloway, Reg. No. 37,850; George W Hoover II, Reg. No. 32,992; Eric S. Hyman, Reg. No. 30,139; William W. Kidd, Reg. No. 31,772; Sang Hui Kim, Reg. No. 40,450; Walter T. Kim, Reg. No. 42,731; Eric T. King, Reg. No. 44,188; Erica W. Kuo, Reg. No. 42,775; George Brian Leavell, Reg. No. 45,436; Kurt P. Leyendecker, Reg. No. 42,799; Gordon R. Lindeen III, Reg. No. 33,192; Jan Carol Little, Reg. No. 41,181; Joseph Lutz, Reg. No. 43,765; Michael J. Mallie, Reg. No. 36,591; Andre L. Marais, under 37 C.F.R. § 10.9(b); Paul A. Mendonsa, Reg. No. 42,879; Clive D. Menezes, Reg. No. 45,493; Chun M. Ng, Reg. No. 36,878; Thien T. Nguyen, Reg. No. 43,835; Thinh V. Nguyen, Reg. No. 42,034; Dennis A. Nicholls, Reg. No. 42,036; Daniel E. Ovanezian, Reg. No. 41,236; Kenneth B. Paley, Reg. No. 38,989; Marina Portnova, Reg. No. P45,750; William F. Ryann, Reg. No. 44,313; James H. Salter, Reg. No. 35,668; William W. Schaal, Reg. No. 39,018; James C. Scheller, Reg. No. 31,195; Jeffrey Sam Smith, Reg. No. 39,377; Maria McCormack Sobrino, Reg. No. 31,639; Stanley W. Sokoloff, Reg. No. 25,128; Judith A. Szepesi, Reg. No. 39,393; Vincent P. Tassinari, Reg. No. 42,179; Edwin H. Taylor, Reg. No. 25,129; John F. Travis, Reg. No. 43,203; Joseph A. Twarowski, Reg. No. 42,191; Tom Van Zandt, Reg. No. 43,219; Lester J. Vincent, Reg. No. 31,460; Glenn E. Von Tersch, Reg. No. 41,364; John Patrick Ward, Reg. No. 40,216; Mark L. Watson, Reg. No. P46,322; Thomas C. Webster, Reg. No. P46,154; Steven D. Yates, Reg. No. 42,242; and Norman Zafman, Reg. No. 26,250; my patent attorneys, and Firasat Ali, Reg. No. 45,715; and Justin M. Dillon, Reg. No. 42,486; my patent agents, of BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP, with offices located at 12400 Wilshire Boulevard, 7th Floor, Los Angeles, California 90025, telephone (310) 207-3800, and James R. Thein, Reg. No. 31,710, my patent attorney with full power of substitution and revocation, to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith.

APPENDIX B

Title 37, Code of Federal Regulations, Section 1.56 Duty to Disclose Information Material to Patentability

(a) A patent by its very nature is affected with a public interest. The public interest is best served, and the most effective patent examination occurs when, at the time an application is being examined, the Office is aware of and evaluates the teachings of all information material to patentability. Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability as defined in this section. The duty to disclosure information exists with respect to each pending claim until the claim is cancelled or withdrawn from consideration, or the application becomes abandoned. Information material to the patentability of a claim that is cancelled or withdrawn from consideration need not be submitted if the information is not material to the patentability of any claim remaining under consideration in the application. There is no duty to submit information which is not material to the patentability of any existing claim. The duty to disclosure all information known to be material to patentability is deemed to be satisfied if all information known to be material to patentability of any claim issued in a patent was cited by the Office or submitted to the Office in the manner prescribed by §§1.97(b)-(d) and 1.98. However, no patent will be granted on an application in connection with which fraud on the Office was practiced or attempted or the duty of disclosure was violated through bad faith or intentional misconduct. The Office encourages applicants to carefully examine:

(1) Prior art cited in search reports of a foreign patent office in a counterpart application, and

(2) The closest information over which individuals associated with the filing or prosecution of a patent application believe any pending claim patentably defines, to make sure that any material information contained therein is disclosed to the Office.

(b) Under this section, information is material to patentability when it is not cumulative to information already of record or being made or record in the application, and

(1) It establishes, by itself or in combination with other information, a prima facie case of unpatentability of a claim; or

(2) It refutes, or is inconsistent with, a position the applicant takes in:

(i) Opposing an argument of unpatentability relied on by the Office, or

(ii) Asserting an argument of patentability.

A prima facie case of unpatentability is established when the information compels a conclusion that a claim is unpatentable under the preponderance of evidence, burden-of-proof standard, giving each term in the claim its broadest reasonable construction consistent with the specification, and before any consideration is given to evidence which may be submitted in an attempt to establish a contrary conclusion of patentability.

(c) Individuals associated with the filing or prosecution of a patent application within the meaning of this section are:

(1) Each inventor named in the application;

(2) Each attorney or agent who prepares or prosecutes the application; and

(3) Every other person who is substantively involved in the preparation or prosecution of the application and who is associated with the inventor, with the assignee or with anyone to whom there is an obligation to assign the application.

(d) Individuals other than the attorney, agent or inventor may comply with this section by disclosing information to the attorney, agent, or inventor.